4L-06

空間 Web 上の m-最近接キーワード検索における代表解の探索 *

西野 良介　津野 貴大　大森 匡　藤田 秀之　新谷 隆彦

電気通信大学大学院　情報理工学研究科 †

1　はじめに

緯度経度情報つきの Web データを用いた情報抽出は最近のデータベース研究の話題の一つであり，その一つに m-最近接キーワード検索 (mCK 検索) がある．これは，Flickr 写真共有データサービスのように，個々の写真データ (オブジェクト) が場所と写真内容に関するタグ複数を持っている状況で，問い合わせとしてキーワード m 個の入力 Q を与えたとき，Q を満たす高々m 個のオブジェクト集合 O のうち，その要素の相互近接度が最も高い集合 O_{opt} を求める問題である [1].

mCK 検索は，オブジェクトの組み合わせを解とするため，上位 k 個の解を列挙すると，地図上の数か所に解が集まりやすいという問題がある．本稿では，Drosou の DisC Diversity[4] の戦略を用い，代表点を使って mCK 検索を行うことで，上位 k 個の解に多様性を持たせる方法を検討する．

2　mCK 検索問題

2.1　mCK 検索問題の定義と解法

本稿では地図上の位置情報を持つ空間 web データをオブジェクトと呼び，各オブジェクトは地図上の 1 点を表すと仮定する．すなわちオブジェクト o は位置情報 $o.l$ とテキスト情報 $o.t$ をもつ．このとき mCK 検索とは，ユーザが m 個のキーワード Q を入力したとき，「オブジェクト o が持つテキスト情報は Q の少なくとも 1 つ以上のキーワードに該当する」という制約下で高々m 個のオブジェクトの集合 $O = \{o_{i1}, o_{i2}, ..., o_{il}\}$ を対象としたとき，Q の各キーワードが少なくとも 1 つの O のオブジェクトによって満たされており，かつ，そのような O のうち，直径 $diam(O)$ が最小となるものを最適解 O_{opt} として返す検索問題である [1]. (ここで，O の直径 $diam(O)$ とは O の 2 オブジェクトの距離 $dist$ の最大値．即ち $diam(O) = \max_{\forall o_i, o_j \in O} dist(o_i, o_j)$.)

mCK 検索の例を示す．図 1 では，地図上にカフェ，学校，池というテキスト情報を持った 3 種類の写真が存在する．図 1 の例で Q = {カフェ，学校，池} で mCK 検索を行うと，3 種類の写真からなるオブジェクト集合で直径が最も小さい集合を探す．すると，赤い線で結ばれたオブジェクト集合が最も直径が小さいのでこれを O_{opt} として返す．O_{opt} が Q に応じた最適な位置となる．

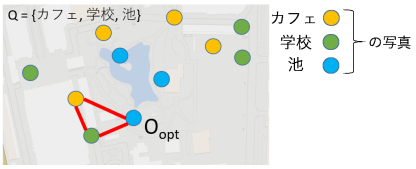

図 1: mCK 検索の例

mCK 検索は m について NP 困難であり [1]，本稿では PE 法 [2][3] を用いる．PE 法では，データベース上で Q のキーワードを 1 つでも満たすようなオブジェクト全てを一つの四分木に格納する．四分木の各ノードには，そのノードに属する点集合の MBR(最小包囲矩形) を持たせている．その後最近接二点探索の戦略を使って mCK 解の直径の候補となる 2 つのオブジェクトペア (o_1, o_2) を小さい順に列挙し，各 (o_1, o_2) が直径となりうる領域の中に残りのキーワードを持つオブジェクト組が存在して mCK 解を構成するかを再帰的に検査する．

2.2　mCK 検索における問題点

図 2 は [3] と同じ Flickr データセット約 24 万件について Q = {sakura，river，temple} で Top-100 mCK 検索を行ったときの mCK 解の位置を東京周辺の地図で表示したものである．地図上の赤点が sakura，緑点が river，青点が temple を表していて，直径が小さい順に青い数字で解を表している．図 2 の 30 番目の解の出現場所を拡大したものが図 3 である．

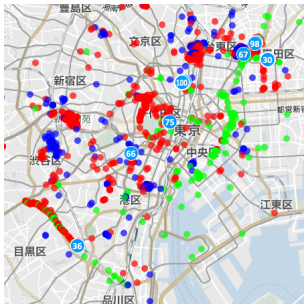

図 2: Q= {sakura，river，temple} Top-100

図 3: 図 2 の 30 番目の解の出現場所の拡大図

図 2 より Q = {sakura，river，temple} で Top-100 mCK 検索を行うと上位 100 解が東京の地図上でどのよ

*Exploring Representative Answers in m-CK Search over Spatial Web

†R.Nishino，T.Tsuno，T.Ohmori，H.Fujita，T.Shintani・U.Electro-Comm.

うに出現しているのかが分かる．30番目の解の出現場所に着目すると，図3のように1か所に解が集まって出現している．mCK検索は，オブジェクトの組み合わせを解とするため，Qのキーワードを持つオブジェクトが密集している場所に直径の候補となるオブジェクトペアが集中してしまい，一か所に解が集中して出現するという問題が発生する．

3　提案手法と予備的実験

代表点を用いたm CK検索：2節で指摘した課題を解決するためにDrosouのDisC Diversityの研究[4]の二次元空間上に分布している点集合について，点集合に支配関係を持たせるため半径rの円内の点をその円の中心点に代表させるという手法をオブジェクト集合に適用してmCK検索を行うことで，解が集中して出現する問題に対応した．手順は以下の通り：

1. 問い合わせQの各キーワードXごとに，データベースからそのキーワードを持つオブジェクトを取得し，キーワード毎のオブジェクト集合$Set[X]$を生成する
2. $Set[X]$の中から，ランダムに1点の代表点oを選び，そのoを中心とした半径rの円内にある他のオブジェクトをoに集約し，oをDataListに格納する（oとoに集約されたオブジェクトは$Set[X]$から削除）
3. $Set[X]$が空になるまで2.の手順を繰り返す．ほかのキーワードについても2~3を行う
4. 選び取った代表点の集合DataListでmCK検索を行う

予備的実験：上記の提案手法によって上位100解の地図上の出現位置の集中が解決するのかを確認するための予備的な評価を行った．データセットは従来手法（図2）と同じものを用いた．解が局所的に集中して出現する問題を解決できたかの確認として，代表点集約の円の半径をr=0.5km，1.0kmとして，Q={sakura，river，temple}で評価実験を行った．図5はr=0.5km，図6はr=1.0kmでTop-100mCK検索を行ったときの東京周辺の地図を拡大したものである．

図4: r=0.5kmの上位100解

図5: r=1.0kmの上位100解

図2，図4，図5の結果より，代表点でmCK検索を行うことで上位100解が散らばって出現し，集約の半径を大きくすると解の出現位置の多様性が増加していることが分かる．

従来手法での解の出現位置の多様性の評価方法の異なる2つの解の間の距離の平均値（diversity）を今回の提案手法の解についても計算し評価を行った．（diversityの数値が大きいほど，それぞれの解が地図上の離れた位置に出現しているといえる．）上位10，20，40，100解における従来手法（noscore）[3]と提案手法（r=0.5km）でのdiversityを計算すると，従来手法のdiversityは2293，9686，18884，22351となり，提案手法では23980，15918，16133，14037となった（diversityの単位はメートル）．Top10，20では提案手法のdiversityが優れているが，それ以降のTop40，Top100で従来手法の方がdiversityの値は大きいことがわかる．これは従来手法の解の出現場所が，地図上の離れた数か所に集中しているためである．

まとめとして，本稿では，Drosouの代表点選出方式をTop-K　mCK検索問題に適用し，特に上位10~20解のとき多様な場所のmCK解の抽出・列挙に有効であることを確認した．多様性尺度の修正や半径rに伴う誤差を入れた順位づけについては本稿発表時に報告したい．

参考文献

[1] T.Guo,X.Cuo,G.Cong”Efficient Algorithms for Answering the m-closest Keywords Query，” ACM SIGMOD, pp.405-418, 2015.

[2] Qiu, Hei, Ohmori, Fujita, ”An Object-Pair Driven Approach for Top-k mCK Query Problem by Using Hilbert R-tree，” 2019 IEEE BigDataSE pp.655-661, 2019

[3] 津野，大森，藤田，新谷 ” 空間 Web データ上の m-最近接キーワード検索問題における点データスコアの導入，” FIT2020 D-005

[4] M.Drosou，E.Pitoura ”DisC Diversity: Result Diversification based on Dissimilarity and Coverage，” VLDB 2013，pp.13-24，2013